

Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function

Georg Mohr, Philip S. Perlman¹ and Alan M. Lambowitz*

Departments of Molecular Genetics, Biochemistry, and Medical Biochemistry, and the Biotechnology Center, The Ohio State University, 484 West 12th Avenue, Columbus, OH 43210 and ¹Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75235-9038, USA

Received July 16, 1993; Revised and Accepted August 23, 1993

ABSTRACT

Many group II introns encode reverse transcriptase-like proteins that potentially function in intron mobility and RNA splicing. We compared 34 intron-encoded open reading frames and four related open reading frames that are not encoded in introns. Many of these open reading frames have a reverse transcriptase-like domain, followed by an additional conserved domain X, and a Zn²⁺-finger-like region. Some open reading frames have lost conserved sequence blocks or key amino acids characteristic of functional reverse transcriptases, and some lack the Zn²⁺-finger-like region. The open reading frames encoded by the chloroplast tRNA^{Lys} genes and the related *Epifagus virginiana matK* open reading frame lack a Zn²⁺-finger-like region and have only remnants of a reverse transcriptase-like domain, but retain a readily identifiable domain X. Several findings lead us to speculate that domain X may function in binding of the intron RNA during reverse transcription and RNA splicing. Overall, our findings are consistent with the hypothesis that all of the known group II intron open reading frames evolved from an ancestral open reading frame, which contained reverse transcriptase, X, and Zn²⁺-finger-like domains, and that the reverse transcriptase and Zn²⁺-finger-like domains were lost in some cases. The retention of domain X in most proteins may reflect an essential function in RNA splicing, which is independent of the reverse transcriptase activity of these proteins.

INTRODUCTION

Group II introns are of interest both because of their RNA-catalyzed splicing mechanism, which resembles that of nuclear pre-mRNA introns, and because they behave as mobile elements (1,2). Group II introns have been found in fungal and plant mitochondria and in chloroplasts (1), and recently, in the

proteobacterium *Azotobacter vinelandii* and the cyanobacterium *Calothrix*, which are related to the probable ancestors of mitochondria and chloroplasts, respectively (3). All group II introns have a conserved secondary structure, which consists of six double helical domains radiating from a central wheel, with the two different structural classes, IIA and IIB, distinguished by specific features (1). The conserved RNA structure catalyzes splicing via formation of a lariat intermediate similar to that formed during the splicing of nuclear pre-mRNA introns (4–6). Although some group II introns self-splice *in vitro*, they require proteins for efficient splicing *in vivo*, presumably to help fold the intron RNA into the catalytically-active structure. Some of these proteins are encoded by chromosomal genes, whereas others, ‘maturases’, are encoded by the introns themselves (7).

The mobility of group II introns has been inferred from their location and distribution in different genes (8) and from the existence of ‘twintrons’ in which one group II intron has integrated into either another group II intron or into a degenerate type of group II intron, referred to as a group III intron (9,10). The first direct evidence for group II intron mobility came from studies showing that two *Saccharomyces cerevisiae* (yeast) group II introns (*cox1* intron 1 and *cox1* intron 2) insert efficiently during crosses into *cox1* alleles lacking these introns (11,12). Similar findings have also been reported for *Kluyveromyces lactis* *cox1* intron 1, a cognate of yeast *cox1* intron 2 (13). All three of these mobile group II introns contain a long open reading frame (ORF), which encodes a reverse transcriptase (RT)-like protein. The proteins encoded by the yeast *cox1* introns 1 and 2 have been shown to be bifunctional: they have an RT activity that may play a role in intron mobility (14), and they also function as maturases in splicing the intron in which they are encoded (15–17).

In addition to yeast *cox1* introns 1 and 2, a number of other group II introns contain ORFs that may encode proteins that function in RNA splicing or intron mobility (8). Some of these ORFs are located in the loop of intron domain IV, whereas others are in-frame with the upstream exon with the bulk of the ORF in the loop of domain IV (1). The yeast *cox1* intron 2 protein,

* To whom correspondence should be addressed at: Department of Molecular Genetics, The Ohio State University, 484 W. Twelfth Avenue, Columbus, OH 43210, USA

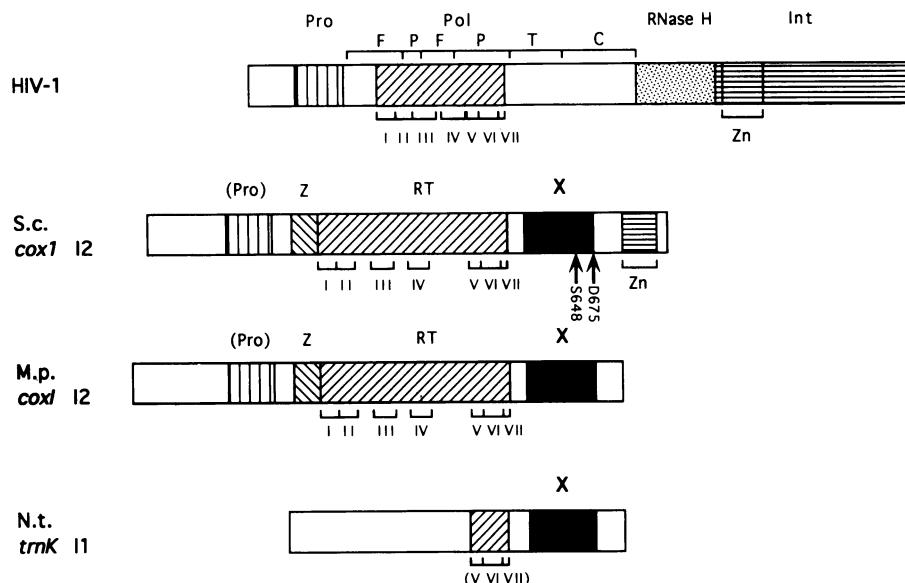


Figure 1. Comparison of protein domains in the HIV-1 *pol* gene and ORFs of group II introns *S.cerevisiae* (yeast) *cox1* intron 2, *M.polymorpha* *cox1* intron 2 and *N.tabacum* *trnK* intron 1. Protein domains are indicated by marked areas. [Pro, protease; (Pro), possible protease domain of intron ORFs; Z, domain Z; Pol and RT, reverse transcriptase domain; X, domain X; Zn, Zn^{2+} -finger-like region, Int, integrase]. Conserved sequence blocks I to VII characteristic of RTs are indicated below the RT domains (18,20); parentheses indicate weak, but recognizable matches for the RT sequence blocks (see Fig. 2). Demarcated regions of the HIV-1 protein indicate structural domains (P, palm; F, fingers; T, thumb; C, connection) identified in the X-ray crystallographic structure of the protein (32). Vertical arrows indicate positions of yeast *cox1* intron 2 mutations, which are associated with the maturase defect in mutant C1082 (Ser648 and Asp675; refs. 14,17). Citations for sequence data are given in Table 1.

shown schematically in Fig. 1, as well as most other group II intron ORFs, have a readily identifiable RT domain. This domain generally includes matches for the seven conserved sequence blocks characteristic of RTs (denoted I to VII; refs. 18–20), although several of the ORFs lack some of the conserved regions or key amino acids characteristic of functional RTs (see below). Phylogenetic comparisons indicate that the RT domains of group II intron-encoded proteins belong to a class characteristic of the LINE1-like or non-long-terminal repeat (non-LTR) family of retroelements. Within this class, the group II intron-encoded proteins comprise a separate subgroup, which is most closely related to the RTs of the *Neurospora* Mauriceville mitochondrial plasmid and bacterial retrons (19–21).

In addition to the RT-like domain, the group II intron ORFs also contain a conserved, upstream domain, Z, which is characteristic of non-LTR retroelements (19,21). Some have a C-terminal Zn^{2+} -finger-like region with similarity to the N-terminal region of retroviral integrase domains (*e.g.*, yeast *cox1* intron 2; refs. 19,21), whereas others lack this region (*e.g.*, *Marchantia polymorpha* *cox1* intron 2; Fig. 1). Finally, some group II intron ORFs also have a weak match to a retroviral protease domain (19,21). A possible function for such a protease is suggested by studies with the yeast *cox1* intron 1 and *cox1* intron 2 proteins, which appear to be synthesized via proteolytic processing of an initial exon-intron fusion protein (15,16,22).

In compiling a summary of group II intron-encoded ORFs for recent reviews (2,8), we noticed that some, exemplified by the tobacco chloroplast tRNA^{Lys} (*trnK*) intron ORF (Fig. 1), lack intact RT and Zn^{2+} -finger-like domains, but retain another conserved domain, denoted X, which is located between the RT and Zn^{2+} -finger-like regions when they are present. Here, we have updated the comparison of group II intron-encoded ORFs and describe the characteristics of domain X.

RESULTS AND DISCUSSION

Sequence comparisons of group II intron-encoded and related ORFs

Computer searches of available protein databases (GenBank, PIR, Swiss Prot), using the BLASTP program with standard parameters (NCBI, BLAST network service) (23), identified 34 intron-encoded ORFs and four other ORFs, which encode related proteins, but are not found in introns (Table 1). Thirty seven of these ORFs were found in either chloroplast DNAs (cpDNAs) or mitochondrial DNAs (mtDNAs), and the remaining ORF is in the group II intron recently reported in the cyanobacterium *Calothrix* spp. These include all known group II intron ORFs currently in the databases. The 17 cpDNA ORFs include nine closely related ORFs in group II introns of the *trnK* genes of different organisms, four in group II introns of protein-encoding genes (*Euglena gracilis* *psbC* intron 2 and *psbD* intron 8, *Scenedesmus obliquus* *petD* intron 1 and *Bryopsis maxima* *rbcL* intron 1), and one in a group III twintron of a protein-encoding gene of *E.gracilis* (*psbC* intron 4; ref. 24, 25). The remaining three ORFs are found in cpDNAs, but are not associated with introns. These are: *Epifagus virginiana* *matK*, which is closely related to the *trnK* intron ORFs; *Astasia longa* ORF456, which is very similar to the *E.gracilis* *psbC* intron 4 ORF; and *E.gracilis* ORF516, which has a region of about 50 amino acids similar to domain X of the *B.maxima* *rbcL* intron 1 ORF. Interestingly, the *E.gracilis* ORF516 is itself interrupted by three group II introns and one group III intron (24).

The twenty ORFs found in mtDNAs include 19 in group II introns and one freestanding ORF (*Marchantia polymorpha* ORF732). Three other RTs or RT-like ORFs, which have been found in mitochondria, do not fall into the group II intron subgroup defined by phylogenetic analysis (20). These are the

Table 1: Characteristics of Group II Intron-Encoded ORFs and Related ORFs^a

Organism ^b	Location ^c	Size ^d (aa)	Domains ^e	Frame ^f	References ^g	Accession No. ^g
Cyanobacterium						
Cthx.	ORF1 1S ^h	584	RT-X-Zn	F	(3)	X71404
Chloroplasts						
S.a.	<i>trnK</i> I1	502 ⁱ	X	F	(35)	X04826
N.t.	<i>trnK</i> I1	509	X	F	(48)	Z00044
S.t.	<i>trnK</i> I1	509	X	F	(49)	Z11741
S.r.	<i>trnK</i> I1	237 ^j	X	F	(50)	JN0302
H.v.	<i>trnK</i> I1	503	X	F	(51,52)	X52765, X07942
O.s.	<i>trnK</i> I1	511 ⁱ	X	F	(53)	X15901
P.c.	<i>trnK</i> I1	515	X	F	(54)	X57097
P.t.	<i>trnK</i> I1	515	X	F	(55)	D11467
M.p.	<i>trnK</i> I1	497 ^k	X	F	(56)	X04465
S.o.	<i>petD</i> I1	608	RT-(X)-Zn	F	(57)	P19593
B.m.	<i>rbcL</i> I1	478 ^l	RT-(X)	F	(58)	X55877
E.g.	<i>psbC</i> I2	466 ^m	X	F	(24)	X70810
E.g.	<i>psbC</i> I4	458	(RT)-(X)	F	(25)	X70810
E.g.	<i>psbD</i> I8	506	(RT)-X ⁿ	F	(24)	X70810
Mitochondria						
T.a.	<i>nad1</i> I4	678	RT-X	F	(59)	X57968
O.b.	<i>nad1</i> I4	671	RT-X	F	(60)	M63034
V.f.	<i>nad1</i> I4	687	RT-X	F	(61)	M30176
M.p.	SrDNA II	501	RT-X	F	(62)	M68929
M.p.	<i>atp9</i> I1	771°	(RT)-X	U	(62)	M68929
M.p.	<i>atpA</i> I1	1065	RT-X-Zn	U	(62)	M68929
M.p.	<i>atpA</i> I2	909	RT-X	U	(62)	M68929
M.p.	<i>cob</i> I3	949	RT-X-Zn	U	(62)	M68929
M.p.	<i>coxl</i> I1	848	(RT)-X-Zn	U	(62)	M68929
M.p.	<i>coxl</i> I2	743	RT-X	U	(62)	M68929
M.p.	<i>coxl</i> I2	744	RT-X	U	(62)	M68929
N.c.	<i>coI</i> I1	835	RT-X-Zn	U	(34)	X14669
P.a.	<i>COI</i> I1	789	RT-X-Zn	U	(63)	X55026
P.a.	<i>COI</i> I4	791	RT-X-Zn	U	(64)	X55026
P.a.	<i>ND5</i> I4	780	(RT)-X	U	(65)	X55026
S.c.	<i>coxl</i> I1	778	RT-X-Zn	U	(66)	M62622
S.c.	<i>coxl</i> I2	786	RT-X-Zn	U	(66)	M62622
K.l.	<i>coxl</i> I1	786	RT-X-Zn	U	(67)	X57546
S.p.	<i>cob</i> I1	807	RT-X-Zn	U	(33)	X54421
Freestanding ORFs						
E.v. (cp)	<i>matK</i>	505 ^p	X	—	(44)	M81884
A.l. (cp)	ORF456	456	(RT)-(X)	—	(30)	X14385
E.g. (cp)	ORF516	516	(X)	—	(24)	X70810
M.p. (mt)	ORF732	732	RT-X	—	(62)	M68929

^a The Table is updated from ref. 8.^b A.l., *Astasia longa*; B.m., *Bryopsis maxima*; Cthx., *Calothrix species*; E.g., *Euglena gracilis*; E.v., *Epifagus virginiana*; H.v., *Hordeum vulgare*; K.l., *Kluveromyces lactis*; M.p., *Marchantia polymorpha*; N.c., *Neurospora crassa*; N.t., *Nicotiana tabacum*; O.b., *Oenothera berteriana*; O.s., *Oryza sativa*; P.a., *Podospora anserina*; P.c., *Pinus contorta*; P.t., *Pinus thunbergii*; S.a., *Sinapis alba*; S.c., *Saccharomyces cerevisiae*; S.o., *Scenedesmus obliquus*; S.p., *Schizosaccharomyces pombe*; S.r., *Secale cereale*; S.t., *Solanum tuberosum*; T.a., *Triticum aestivum*; V.f., *Vicia faba*.^c Gene designations follow those in use for each organism, e.g., *ND* and *nad* both refer to NADH dehydrogenase subunit genes.^d Size in amino acids estimated from the ORF sequence. In case of fusion to the 5' exon, the N-terminus is assumed to correspond to the first amino acid encoded within the intron.^e Domains: RT, reverse transcriptase; (RT), RT domain lacking conserved YXDD sequence; X, domain X; (X), domain X having less than 40% match to the overall consensus sequence (Fig. 4); Zn, Zn²⁺-finger-like region.^f Frame: F, freestanding in intron; U, fused to upstream exon.^g Source of sequence data used here.^h Gene has not been identified. Intron number is from database file (X71404).ⁱ The lengths of these ORFs was corrected in a subsequent study (J. Palmer, personal communication). Published sizes are 524 (S.a.) and 542 (O.s.) amino acids.^j C-terminal segment of intron ORF only.^k Published size is 370 amino acids; a single frameshift adds 127 amino acids similar to other *trnK* ORFs.^l Published size is 274 amino acids; three frameshifts in conserved regions of the RT domain add 204 amino acids having similarity to RTs.^m Published size is 241 amino acids; an additional 225 amino acids homologous to other RTs are in frame upstream of the first Met codon of the ORF.ⁿ A C₂H₂-type Zn²⁺-finger-like domain has been identified in this ORF (24), but it overlaps the C-terminal part of the RT domain and the N-terminal part of domain X and is dissimilar to the other intron ORF Zn²⁺-finger-like domains.^o Published size is 681 amino acids; a single frameshift in domain X replaces 34 C-terminal amino acids with 124 amino acids similar to other domain X sequences.^p Published size is 439 amino acids; a single frameshift adds 66 amino acids similar to *trnK* ORFs to the N-terminus of the putative protein.

RT encoded by the *Neurospora crassa* Mauriceville mitochondrial plasmid (26) and the RT-like ORFs in *Chlamydomonas reinhardtii* and *Oenothera berteriana* mtDNAs (27,28).

As summarized in Table 1, most of the group II intron-encoded ORFs contain a domain having similarity to RTs, followed by a conserved domain X, defined below, and some also contain a C-terminal Zn²⁺-finger-like region. The cyanobacterial ORF contains all three of these domains and is most closely related to the ORF of the alga *B. maxima rbcL* intron 1. In most cases, the RT-like domain contains readily identifiable matches for all seven conserved sequence blocks characteristic of functional RTs. However, several of the ORFs (*E. gracilis psbC* intron 2 and *psbD* intron 8; *M. polymorpha coxI* intron 1 and *atp9* intron 1; and *P. anserina ND5* intron 4) lack the conserved YXDD sequence, which has been shown to be required for the activity of the retrovirus HIV-1 RT (29). In addition, the ORFs in group II introns of the *nad1* genes of plant mtDNAs contain the YXDD sequence, but are missing conserved sequence block I.

Two other cpDNA ORFs, *E. gracilis psbC* intron 4 and *A. longa* ORF456, had been reported to be related to each other, but unrelated to other group II intron ORFs (25,30). As shown in Fig. 2, however, both these ORFs have detectable similarity to the RT domains of group II intron ORFs, although they lack the highly conserved YXDD sequence and other amino acids characteristic of functional RTs (e.g., box I 'K' and box IV 'PQQ'). It seems likely that these ORFs originally encoded RT-like proteins, which have diverged from the conserved RT sequence and lost RT activity. The domain X sequences in these proteins have also diverged substantially (see below).

All the chloroplast *trnK* intron ORFs and the *matK* ORF of *E. virginiana* cpDNA are closely related to each other throughout their lengths, but have little similarity to other group II intron proteins, except for one conserved domain, X (see below). A clue to the evolutionary origin of the *trnK* proteins came from sequence alignments, which showed that the region of these proteins immediately upstream of domain X has some similarity to conserved sequence blocks V, VI and VII of the RT-domains of other group II intron ORFs (Fig. 3). However, this similarity does not extend upstream of conserved sequence block V, and the region of the *trnK* ORFs corresponding to sequence block V lacks the conserved YXDD sequence. Together, these findings suggest that the *trnK* and *matK* ORF proteins also evolved from an RT-like group II intron-encoded protein. The *trnK* ORFs have retained domain X, but their RT domains have mostly diverged. The finding that the C-terminal region of the RT-domain is still recognizable may reflect its proximity to and participation in some function carried out by domain X.

Characteristics of domain X

Domain X is defined by the sequence alignments shown in Fig. 4. The domain spans about 100 amino acids beginning downstream of the last conserved sequence block of the RT domain. This location corresponds to the 'tether' region between the polymerase and RNase H domains of retroviral *pol* genes (31) and to the 'thumb' and 'connection' domains identified in the X-ray crystallographic structure of the HIV-1 RT (32) (Fig. 1). Other workers had noted previously that the similarity between fungal group II intron ORFs extended beyond sequence block VII of the RT domain (33,34) and that the mustard *trnK* ORF has a region similar to the fungal group II intron-encoded ORFs (35). To maximize the similarity of domain X sequences, it was necessary to insert a gap of variable length separating the N-

B.m. <i>rbcL</i> 11	MAQVRNLKYRIFMAKLEGNRRTLRLRQLK	29
S.o. <i>petD</i> 11	MKRKIKLKYRKLRKQOFIKVKEPNIGIDMVECHRTLENKQESIAKACREGNIVLVEKLAQE	60
A.1. ORF456	MISFSLLINPSNLSWDKVTNLIFLKQKRLPKVSYYVDKRLKLYELQKL	47
E.g. <i>psbC</i> 14	MIFYEIFQVSFYQNFLWEKSVRLLFKVQRRLFKVSYIHDKKKLYELQKL	49

B.m. <i>rbcL</i> 11	MLDSSRHKKILLSVRRRTVINDGKIIPIGDVKLIKDKNDRMVLVEKLYKM--NIQYEPCKPV	87
S.o. <i>petD</i> 11	IVNSNSFGRAVAVGTVW--SSKGSRSPGLSRESFKTNKNVAMMATEQITSMSPHVKATPL	119
A.1. ORF456	IVQLNYSRLLAIIKLVNHSVFNENLPGDGVVSLNFYTESFELNEFLKY--NWNNWIFQNL	104
E.g. <i>psbC</i> 14	IFQSLSNCARLLAIERTVQSLNSKKKISVGDKTTLNFLERFELNEYLRK--NWNNWPKQSL	106

I	II	
-hh-h-K>	<h>-R----->	
B.m. <i>rbcL</i> 11	KKIQIPKRSNGKTRPLGIPTILDRCIQAIIVLNALQPEW&CPEANSYGPGRGRSAHDAGIR	147
S.o. <i>petD</i> 11	SRIVIPKRDGSARPLSIPSYSTDRCILQALYKLAIEPMAEVAEADLSSYGPGRPMRNVSWAVGR	179
A.1. ORF456	KKVSLEFDNDGKIIIVKVKVFPV18DRWCVYLVKVFAIEPVWHLFHPPNLGYRSQFYIIEQEL	164
E.g. <i>psbC</i> 14	RKQKVFDLNENLISDTISDRS9QVLIKFALAEPVWHEAFPHPFNQGFRYVNPVYKVQQA	166

III	A	
-hh-hDh-h-K>	<h>-Go->	
B.m. <i>rbcL</i> 11	VFNVLSVKSGNNGNNKWPWLDAHIEGFDFDICHQDILEKLDNFPAKKLINRNLKAGYME--	206
S.o. <i>petD</i> 11	VLNGLANNP-LANQYQVWEIDIKGCDVNINHOFI--SQVTFPIPKKILWALKCGYIERN	235
A.1. ORF456	ILLNLNSKSFSGSKKRVLLKDLSQNCINNYSFIME-KLIAPRC1KLGIFKLLKEGFLV--	221
E.g. <i>psbC</i> 14	ILLNLNSIISFGSKKRIKLVELNCNFSIINYDYLWK-KLIAPRN1KLGIFRLLGEFLGPNL--	223

IV	V	
<h>POG--pP-hh--h>	<h>-DPhhh>	
B.m. <i>rbcL</i> 11	EDIFSKTEKPKGSII^bPLLLAIALYEEFLKASPDSTGRVRGRNVVRYADDIFVIC	265
S.o. <i>petD</i> 11	SNTLOPPTTGTGVPGGI^bPLLMH^bTLDDGLEYHKIKQKSSQSKGKNTYCRYADDWLT	295
A.1. ORF456	---EFSNNC1IFNKVDFBSLLNLLNIFLNGIEKL-----HNCIRGYGFLLF	263
E.g. <i>psbC</i> 14	---HFPENECQIST-FSSLLNVMNLNGENV-----HNCVRGYMMFLF	264

VI	VII	
<Oh-h-cK--h>	<h>-Go->	
B.m. <i>rbcL</i> 11	ASKTEAQQTKKQISGWLK^bKLAFAPNPKI--SDIDKG^bPLGVDRVSTNASHKT	318
S.o. <i>petD</i> 11	TTEETALIPLKFLAVRGLEVKLAKTTKNIINDRQ^bPLSFRPKVVRRNRKLT	351
A.1. ORF456	NPIDNEKELLSKLYFLSKLDDLFNISEIEL---SSIINGDFPLGHWFHKFSYSYNLCI	316
E.g. <i>psbC</i> 14	RPMDEKILANQILSLLYTRGKIKNNSKFL---VSNTKG^bPLGHWFHKFSEKVKNGLS	317

Figure 2. Sequence alignment of RT domains of algal ORFs. Sequences were aligned with the CLUSTAL program (45,46), and the alignment was manually refined to match those of other group II intron-encoded ORFs (20). Amino acids conserved in reverse transcriptases are indicated above the alignments (b, basic; c, charged; h, hydrophobic; o, aromatic; p, polar) (20). Numbers at the right of the ORF sequences indicate amino acid positions. Bold letters indicate conserved or similar amino acids. Similar amino acids are defined as D and E; R and K; F and Y; L, I, V and M; S and T. Citations for sequence data are given in Table 1.

terminal part from the remainder, as well as a few short gaps (1 to 4 amino acids) elsewhere. The most strongly conserved region of domain X includes the consensus sequence SX₃₋₆TLAXKXK. Most of the domain X sequences have a large excess of basic over acidic amino acids (Arg+Lys: 13 to 22; Asp+Glu: 3 to 13), consistent with a role in nucleic acid binding. Exceptions, however, are the *trnK* intron ORFs of *Pinus contorta* and *Pinus thunbergii*, which have roughly equal numbers of positively and negatively charged amino acids (Arg+Lys/Asp+Glu: 13/13 and 15/13, respectively).

Secondary structure predictions for the domain Xs of seven different ORFs using the PredictProtein service (EMBL Heidelberg; (36)) consistently identified one potential α -helical region (α -A) and one potential β -sheet region (β -A), both having somewhat variable boundaries in different ORFs (Fig. 4). The program predicted two other potential β -sheet regions (β -B and β -C) in five or more proteins, but gave no consistent structural predictions for other regions.

A 100 amino acid region of the chloroplast *trnK* intron ORFs, which we now define to be part of domain X, was reported to have amino acid sequence similarity to a putative 17 kDa nucleic acid binding protein of barley stripe mosaic hordeivirus (37). It was noted further that the region of similarity corresponds to the 'tether' domain of retroviral RTs. Otherwise, searches of all available protein databases using the BLASTP program (23) failed to detect homologs of domain X in other proteins, including other RTs or nucleic acid-binding proteins. Likewise, analysis of domain X sequences failed to detect good matches to known RNA binding motifs (Zn²⁺-finger, Zn²⁺-knuckle, arginine-rich motif, RNP motif/RRM, RGG box, KH motif, dsRNA-binding motif, helix-turn-helix motif; see 38,39 and references therein). The

	V	VI	VII
	<h--o-DDhh-->	<----Gh-h--cK--h>	<----hLG--h>
Chloroplasts			
S.a. <i>trnK</i> 11	291-DPLFLHYVRHYHGKYIYLATKDTPLM-MNKWKYVFVNWLQCYFVNFQSQKINI---NQLSKDNFDPDLYGFSSLRLN-359		
N.t. <i>trnK</i> 11	276-DPFMHYVRVYQCKSILASKGTFLL-MNKWKYFLVNFWQOCHSLCSFCHTGRHI---NOLNSNHSRDPMDGYLSVRLN-344		
S.t. <i>trnK</i> 11	276-DPFMHYVRVYEGKSILASKGTFPL-MNKWKYFLVNFWQOCHSMFYFHTGRHI---NOLNSNHSRDPMDGYLSVRLN-344		
S.r. <i>trnK</i> 11	1-DPLHYVRVYQGKAILASKGTFPL-KKKWKCYLINFQYYFCFTWPTPRRIHI---NQLANSCDFDPDGYLSVPKS-343		
H.v. <i>trnK</i> 11	275-DPLHYVRVYQGKAILASKGFSFL-KKKWKCYLINFWFQYYFFWTPTPRRIHI---NQLANSCDFDPDGYLSVPKS-343		
O.s. <i>trnK</i> 11	276-DPLHYVRVYQGKAILASKGTL---KKKKWCYLVNLWQYOSFSFTWPQPRRIHI---NQLANSCDFDPDGYLSVPKS-343		
P.c. <i>trnK</i> 11	270-DPKHYVRVYGRPIIAKGAHLL-VKKKRCYLLIFRQFYFHLSWESEPYRVC5---HQLSKNCSSSPGYFLVRVMN-338		
P.t. <i>trnK</i> 11	270-DPKHYVRVYGRPIIAKGADLL-VKKKRCYLLIFRQFYFHLSWESEPYRVC5---HQLSKNCSSSPGYFLVRVMN-338		
M.p. <i>trnK</i> 11	258-ISHSYVYRKNNLITLNDRNLLNLEWNRDFLFLPQKVFNVWFKSRSLI---ONFYKNSFSPGVYMFRIESQ-327		
E.v. matK	274-EPLCHYFRYRRQYIILASKGTSFL-LNKWKCYVITFQWYFSLWFFSRSIYI---KKLNNNSFEILTYHSSLNIN-342		
Mitochondria			
T.a. nad1 I4	318-LIRICYARYADDLLGIVGAFL-LIEIQKRITHFLQSGNLNWVGSAGSTT---IAARSTVEPPGTIVIREVPP-386		
O.b. nad1 I4	312-FIRICYARYADDLLGIVGAFL-LIEIQKRIAHFLQSGNLNWVGSAGSTT---IAARSTVEPPGTIVIREVPP-380		
V.f. nad1 I4	314-FIRICYARYADDLLGIVGAFL-LIEIQKRIAHFLQSGNLNWVGSAGSTT---IAAWSTVEPPGMIVIEWPP-382		
M.p. SrDNA I12	269-GSRVRYVRADDWVIGWPGKAL-AVQIEKEVSTFLQEKLKLQLQAETKTRI---TNLRSREALLGMEIKVVPs-338		
M.p. atp9 I1	536-LKLKLYLRAVDNMILGWVGMSQD-ALDIRKAQAVNQFLQEELKMLDNQEOKS1---LHKASEMAKYLALGVMMYGT-605		
M.p. atpA I1	593-PRNITYVYRVAFFIVGPGDKAF-AKDATTSISNFKISSHEFEVSYEMH---TKSGKTPPLFLGYL1RDSK-561		
M.p. atpA 12	486-NIKRIVYRVAADDLFLVIEGKPLD-SKTIRGEINDFLQSAHLALHEMKRDLNIV---HARSDWVRPMGFDIDQARMT-553		
M.p. cob I3	472-FVRYIACRYADDLMLAVSGSKAL-AREVMERVSFLRFLKDVLMHNPEPCTRL---GHVNEEAKATLGNRLLGKPN-541		
M.p. coxI 1	467-YARAFVYRVAAGNFGFLGIAPREL-VATVKSRIQVONSELNEHLELTGGSMS---HMAESVSKFLGFLVSGGL-534		
M.p. coxI 2	513-FRRILYVYRVAADDFLVILSGSQTLE-TFAIQASLONFLHRSLRSLLESLEKTTVV---SHLANKGFLHFLGTYCKRTRs-522		
M.p. coxII 12	499-YRMRMVEYRRYGGDFLMARMRGFLPSD-AKAIRDECTTFLREKLKLLNMRKTHM---KHMWSGVPFLGPHRIGRVRV-567		
N.c. coxI II	479-TVRVNYVYRVAADDIIFVGEWSKH-TAAVILEVKQSFTVNQGLRFLNPDRTG1---TKYSVDFVKFLGYKMGKAPHM-548		
P.o. COX II	463-LAEEMYVYRVAADDWLWLGIVGWSSET-ARAKERMAAYLKDMLKLEMSEKTRLM---TNASEDCKAYFLGTEIOQRMS-532		
P.a. COX 1	465-GIRVRYVYRVAADDWVIGIQQDEL-VAKIKECKAFLRDLKLESEKTR1---TNITEKEVRFLGVDDIKRDS-534		
P.a. ND5 I4	521-FKKVLYVYRANVWVIGWVGSKED-CNLLEKEIIFKKFDKLNGIAGASSRASS [11] TANQAKVFLPDLTVFWSH-601		
S.c. coxI 1	457-FKRYAVYRVAADDLILWGWLGSND-CKNIK1RDLNFL-NSLGLTINEEKT1---TCATCEPATRPLGYNISI1TPL-525		
S.c. coxI 2	481-FKRAYVYRVAADDIILWGWMGSND-CKNIK1NDINFLNPKENLGMNSINIDKSV1---KHSKEGVSPFLGYDVKVTWP-549		
K.l. coxI 11	481-FKACFVYRVAADDIIGWMGSHD-CKNIK1NDMNFLKENLGMNSINMDK5II---KHSKEGVSPFLGYDVKVTWP-549		
S.p. cob II	503-SNKLWVYRVAADDWIVAVNGSYQ-TKEILAKITCFCS-SIIGLTVSPTRTK1---TNSYTDKILFLGNTNISHSKN-571		
Consensus	o bh YVRYADD hhGh Gp	hb h Fh h L h akp h	h FLG h

Figure 3. Sequence alignments of *trnK* and *matK* ORFs with the region containing conserved blocks V, VI and VII of the RT domains of other group II intron ORFs. Sequence alignments for the fungal group II intron-encoded ORFs are from ref. 20. Additional sequences were added sequentially, and the alignments were refined manually to maximize similarity. Conserved amino acids in sequence blocks V to VII are indicated above the alignment (20). Numbers flanking the ORF sequences indicate amino acid positions. Amino acids found in the majority of group II intron-encoded ORFs and related ORFs are indicated in bold and summarized below the alignment (a, acidic; b, basic; h, hydrophobic; o, aromatic; p, polar). Similar amino acids are as defined in Fig. 2. Citations for sequence data are given in Table 1.

RTs encoded by the *N. crassa* Mauriceville mitochondrial plasmid and the *E. coli* msDNA retrons, representing the two closest relatives of the group II intron RTs (20), had only 28 and 30% matches to the overall consensus sequence for the domain X regions of the group II intron ORFs (Fig. 4). Thus, strong conservation of domain X sequences appears to be restricted to the group II intron and related proteins, possibly reflecting a specific interaction with group II intron RNAs.

Phylogenetic analysis using the PAUP program (40) indicates that the domain X sequences of the *trnK* introns and the *matK* ORF form a separate class, which has diverged from those of the mtDNA ORFs. Among the other ORFs, those that deviate most in the domain X region are the cyanobacterial ORF and a number of the algal chloroplast ORFs (35–41% match to the overall consensus sequence, compared to 54–72% for the cpDNA *trnK* ORFs and 76–93% for the fungal mtDNA intron ORFs; Fig. 4). The deviations are particularly evident in the SX_{3–6}TAXXKX region, which is strongly conserved in the other ORFs. The *E.gracilis psbC* intron 4 ORF and *A.longa* ORF456, whose RT domains deviate from the RT consensus sequences, have also diverged in domain X (35–37% match to the overall consensus sequence). Interestingly, the domain X sequences of the *E.gracilis psbC* intron 2 ORF and some of the *Marchantia* mtDNA intron ORFs appear to be more closely related to fungal mtDNA sequences than to other domain X sequences in *Euglena* cpDNA or *Marchantia* mtDNA, respectively. The domain X sequence of the cyanobacterial ORF is most closely related to that of *B.maxima rbcL* intron 1 (31% identity and 44% similarity over the entire domain X sequence).

Possible functions of domain X

Although domain X has little detectable amino acid sequence similarity to retroviral RTs, its location corresponds to the HIV-1 RT ‘thumb’ and ‘connection’ domains, which have been

suggested to function in binding of the template RNA and tRNA primer (32). By analogy, we speculate that domain X may function in the specific binding of the intron RNA for reverse transcription and RNA splicing and possibly in binding any nucleic acid primer used by the RT.

We found previously that the RT encoded by the yeast group II intron *cox1* intron 2 initiates cDNA synthesis near the 3' end of the intron and immediately downstream in exon 3, and we suggested that binding of the intron-encoded protein at those locations could also be related to its function in RNA splicing (14). If domain X is involved in binding the intron RNA, mutations in this domain should affect the ability of the RT to initiate cDNA synthesis specifically at the 3' end of the intron, as well as splicing activity. Support for the latter prediction comes from analysis of yeast mutant C1082, which has six missense mutations in the *cox1* intron 2 ORF (14,17). This mutant is deficient in both RT and splicing activity, and splicing activity has been shown to be increased in recombinant strains, which correct only the two mutations in domain X, one at Ser648, which is conserved in 30/38 sequences, and the other at Asp675, which is not strongly conserved (see Fig. 4; 14,17). These findings suggest that Ser648 may be a critical amino acid required for RNA binding.

Evolutionary relationships of group II intron-encoded proteins

As noted previously, ORFs are found in only a minority of group II introns, all of which belong to subgroup IIA, and the ORFs are always inserted in domain IV, a substructure that does not play a major role in splicing (1,41,42). These findings suggest that ORFs were inserted into domain IV of pre-existing group II introns and that continuity with the upstream exon and maturase activity were secondary adaptations in some cases (8,43). It seems likely that the group II intron-encoded proteins were initially RTs that functioned to mobilize the introns and that some of these

				Match (%)
Cyanobacterium				
Cthx ORF584	357-YKTLIKPSVKSVKKHYDDIAR-----	IIDNHKNAQETLISKLNPV-		45
Chloroplasts				39
S.a. <i>trnK</i> II	380-KLDNSNIQISSIIGSLAKDKFC-----	NVLGHPIISKATWMODSSDSDILNRFVRI-		
N.t. <i>trnK</i> II	365-KFDTIVPPIPLIGSLAKANFC-----	TVLGHPIISKLPVNSLSDSDIDRGRGI-		
S.t. <i>trnK</i> II	365-KFETLVPIPLIGSLAKAHFC-----	TVLGHPIISKWPNSLSDSDIDRGRGI-		
S.r. <i>trnK</i> II	91-KFDTIVPATPLIGLQLKAQFC-----	TGSGHPIISKPIWTDLSWDIDRGRGI-		
H.v. <i>trnK</i> II	364-KFDTIVPATPLIGLQLKAQFC-----	TGSGHPIISKPIWTDLSWDIDRGRGI-		
O.s. <i>trnK</i> II	365-ENSFLIDITPLIGSLAKAOFC-----	TGSGHPIISKPIWTDLSWDIDRGRGI-		
P.c. <i>trnK</i> II	359-EIDPIVPPIPLIGLQLATEKFC-----	DISGSPPISKLWSLSDIDRGRGI-----		
P.t. <i>trnK</i> II	365-EDMPIVPPIPLIGLQLATEKFC-----	DISGSPPISKLWSLSDIDRGRGI-----		
M.p. <i>trnK</i> II	349-EFCSIIPLVPLIRLAKKEKFC-----	DVLGRPLCKLWSLSDIDRGRGI-----		
S.o. <i>petD</i> II	355-TSQCWIPISA-EKNPFRKNIKA-----	ISKTRKSLDTEINAV-----		
B.m. <i>rbcL</i> II	325-KMILLIREPSTKAIQDITQQLKK-----	EWTPLPRSLIEVTUFLRNPM-----		
E.g. <i>psbC</i> II	325-FFKFPVTFKVFPLKRLRLOP-----	HPFKFPRIGNVRLLFEDPFLILRNPM-----		
E.g. <i>psbC</i> II	317-NGISAIPLSLNNYQOFPLNHNVRK-----	IVNNSNYGSVVKASKLYPV-----		
E.g. <i>psbD</i> II	342-YKFVLSPILEKIVKVKKVLNL-----	CKKFSKPLVLFNDNPKLPMKSWFNS-----		
Mitochondria				
T.a. <i>nad1</i> I4	537-PIQEAPPIKIKILRLRDRGLI-----	-SRRGPPIIHVASLTINVSNDIIRVNWAGI-----		
O.b. <i>nad1</i> I4	531-PIRERAPPIKIKILRLRDRGLI-----	-SRRGPPIIHVACLTINSDGDIWVNWAGI-----		
V.f. <i>nad1</i> I4	533-PIQEAPPIKIKILRLRDRGII-----	-SRRGPPIIHVACLTINSDGDIWVNWAGI-----		
M.p. <i>srDNA</i> II1	357-RMRUCIPIDILIGRLQMGAC-----	-DEKGTPKAVTKWVQEMGEMINKYMAV-----		
M.p. <i>atp9</i> II	626-RQMLMAPNIDMLTALEQLOGA-----	-KNNAKGLARATTSNRVSSGDKHINWVSS-----		
M.p. <i>atpA</i> II	791-TVYLEMPYSKREKRLRSVGF-----	-RDEQGAPFKSLPQTELEDIQMDIFYKOK-----		
M.p. <i>atpA</i> II	742-NLYMRCPSMHRECLRSNIV-----	-AKNTRPPIAVTTSNLDLTIIDWPKSK-----		
M.p. <i>cob</i> II	705-RMOMYAPLRSRMLDRLARGII-----	-NAEGRPTSVPLAATDQDVTITQJGGSV-----		
M.p. <i>coxi</i> II	687-PLOQILAPLKEMRKLKLMGLI-----	-TEENKPCCVVRLQIOLDEDIVLNWNSV-----		
M.p. <i>coxi</i> II	604-RLRVCAPLTKLFLYKLGKFGV-----	-KRNEGMGVYVPTARRNNTPLHDADILELYNK-----		
M.p. <i>coxi</i> II	588-IILTMADMMQNLQMLQOQAYL-----	-AGNCQDPLPNGLMSLQSEANRNSMS-----		
N.C. <i>cob</i> II	598-IRERKPTDPLTQMLQMLQ-----	-KNDLNDHADIXINYNSVG-----		
P.A. <i>cob</i> II	554-STVNAPISKLVTKLADQIVIWKSALNEDWLIPQILKVNLPPLDIIYLRYKMI-----			
P.A. <i>cob</i> II	557-RILYFVTPRDINNKLQKSIYKTYTSANGRKLAQPAINTKIFLDHRSILQDLYRNVAN-----			
P.A. <i>NDS</i> II4	412-ERIELEAPKQIQLQNEAGFI-----	-TNGKPSPKLWQHMYNDTITLTYNSV-----		
S.C. <i>coxi</i> II	547-RPILQAPLTKNKLQNLQNLQNLQ-----	-KHNNGRMGRVFTTRVQFLVNLNKK-----		
S.C. <i>coxi</i> II	571-SLVVNAPIRSITVMQNLKHKNGYC-----	-SHGILQKPRGVRGLRHEEMKTTILMVLAV-----		
K.I. <i>coxi</i> II	571-SLVVNAPIRSITVMQNLKHKNGYC-----	-SCQVIGKPRGVRGLRHEEMKTTILMVLAV-----		
S.p. <i>cob</i> II	587-FILLSMAPDRAKMLQREYGLM-----	-LNNHGRSVLRWLVRDPOIIGLANSI-----		
Freestanding ORFs				
E.V. <i>matK</i>	363-KVNILVPIIILSIS--DKEKFF-----	-NVLGHPTSNTLNFVWDLSDYNIIGRLECI-----		
A.I. <i>ORF456</i>	315-NNLCPFSPSNMNFNLTRIKV-----	-IINNSNYGSVKKHYDDIAR-----		
E.g. <i>orf516</i>	300-SNILLHINKHKOPIFYKLSLK-----	-IVQMLFLKSVFPLVNLNKK-----		
M.p. <i>ORF732</i>	548-GLSLWDMQVKMNRLAEGKFC-----	-DKSGHPPKPNFAYQFQPSYSVARMASI-----		
Mitochondrial Plasmid				
N.C. <i>Maur.</i>	467-EFTAWSPLKRMGFFFSPAGP-----	-TILSSIFNNNSTLNLQEPSPDSLRLYRKGWV-----		
Bacterium				
E.C. <i>msDNA</i>	390-ETSYPELFREKTDSKKKEINL-----	-NIFKSNKETKYLFDLSDGCTADLKKVPERV-----		
Structure		...<-B-A->...<->A-A-----		
Consensus alga	hh P I cb hKh	h hh h h		
Consensus <i>trnK</i>	K DphhPhG LAK FC	h GHPISK WpLDSL DiHDF I		
Consensus <i>mt</i>	h h APh hh bL Goh	P h hh ha Ih o Sh		
Consensus all	h h Ph hh BL G h	P h h Ih o h		
Structure		...<->B-B->...<->B-C->-----*		*
Consensus alga	hb NY o h h oh b o	K b h		
Consensus <i>trnK</i>	CRNlPHYYSGSSKKKTL RYHKYIILRLSCA	bTLARKHH STHRTPhKbLGS LLE		
Consensus <i>mt</i>	RGHANYY FA N h Yhrh Sch	TLA K Kl ph Khb KogDdL		
Consensus all	bghhNYy o N h h Yhhb S	TLA K K p bhh boG L		

Figure 4. Alignment of domain X sequences of group II intron-encoded and related ORFs. Fungal mtDNA sequences were aligned by using the CLUSTAL program (45,46), and additional sequences were added by using the results of database searches with BLASTP (23) as a guide and then manually refined. Numbers flanking the ORF sequences indicate amino acid positions. Conserved amino acids are indicated in bold. * indicates the positions of yeast *coxi* I intron 2 mutations, which are associated with splicing defect in mutant C1082 (Ser648 and Asp675; refs. 14,17). Horizontal arrows indicate potential secondary structure elements predicted by PredictProtein Service (36). α -A, α -helix; β -A, β -B, β -C, β -sheet; solid lines demarcate core region of secondary structure elements predicted in seven ORFs (S.c. *coxi* I and 2; S.p. *cob* intron 1; N.c. *cob* intron 1; P.a. *NDS* intron 4; M.p. *trnK* intron 1; and S.t. *trnK* intron 1); dotted lines demarcate extensions of the core region predicted in five or more of these ORFs. Three consensus sequences were generated for subsets of the ORFs: 'alga', algal chloroplast ORFs, excluding the *trnK* and *matK* ORFs; '*trnK*', nine *trnK* ORFs plus the *matK* ORF; 'mt', 20 mitochondrial ORFs. 'all', consensus sequence for all ORFs. The consensus sequences identify amino acids that are identical or similar in at least 50% of the sequences (a, acidic; b, basic; h, hydrophobic; o, aromatic; p, polar). Similar amino acids are defined as in Fig. 2. In formulating the overall consensus sequence, similar sequences (*trnK* introns and *matK*, *plant* *nad1* introns, S.c. *coxi* I intron 2 and K.I. *coxi* I) were treated as single entry to minimize bias. The Maurieville plasmid sequence is from ref. 26 and the *E. coli* msDNA retron sequence is from ref. 47. Other citations for sequence data are as in Table 1. The domain X sequences used in the alignments incorporate 'revisions' indicated in the legend of Table 1.

proteins adapted to function in RNA splicing by virtue of their ability to bind specifically to the intron RNA (8,14,43). With time, the introns may have lost the ability to self-splice and became dependent on the intron-encoded maturase to help them fold into the catalytically-active structure. Given the relationship among all group II intron-encoded ORFs found here, it is possible that the insertion of the ORF occurred only once, after the divergence of group IIA and IIB introns, giving rise to a mobile element, which dispersed to other locations. The presence of group II introns having related ORFs in both mitochondria and chloroplasts could reflect either horizontal transfer of a mobile intron or ORF between organelles or that the ORF insertion occurred in an ancient bacterium, prior to the evolution of these organelles. The presence of group II introns encoding similar ORFs in bacterial species, which are related to the probable ancestors of mitochondria and chloroplasts (3), is consistent with the second possibility.

Our findings strongly suggest that an ancestral group II intron ORF contained RT, X, and Zn^{2+} -finger-like domains and that the RT and Zn^{2+} -finger-like domains diverged or were lost in

some cases. The conservation of domain X and the C-terminal part of the RT domain in the *trnK* intron ORFs is readily accounted for by the hypothesis that they have an essential function in RNA splicing, which is independent of RT activity. There appears to be no fundamental requirement that group II intron-encoded proteins retain RT activity and indeed there may be selection against this activity, if it is deleterious to the host. Previous work showed that the yeast *coxi* I protein is an active RT in some strains, but inactive in others, presumably due to missense mutations in the *coxi* I ORF in those strains (14). Significantly, the yeast *coxi* I protein, which has lost RT activity, remains functional as a maturase, as expected if splicing activity requires only that the protein retain its ability to bind specifically to the intron RNA (14).

The chloroplast genome of the nonphotosynthetic parasitic plant *E. virginiana* has lost non-essential genes related to photosynthesis, but retains a freestanding *matK* ORF, which is closely related to the *trnK* intron ORFs (44). It was suggested that retention of the *matK* ORF reflects an essential function in the splicing of several group II introns that are retained in the

E.virginiana chloroplast genome (44), and our finding that the *matK* ORF contains a conserved domain X is consistent with this suggestion.

ACKNOWLEDGEMENTS

We thank Dr Richard Hallick (U. Arizona) for the communication of unpublished data. This work was supported by NIH grants GM37949 to A.M.L. and GM31480 to P.S.P.

REFERENCES

- Michel, F., Umesono, K. and Ozeki, H. (1989) *Gene* **82**, 5–30.
- Saldanha, R., Mohr, G., Belfort, M. and Lambowitz, A.M. (1993) *FASEB J.* **7**, 15–24.
- Ferat, J.-L., and Michel, F. (1993) *Nature* **364**, 358–361.
- Peebles, C.L., Perlman, P.S., Mecklenburg, K.L., Petrillo, M.L., Tabor, J.H., Jarrell, K.A. and Cheng, H.-L. (1986) *Cell* **44**, 213–223.
- Van der Veen, R., Arnberg, A.C., van der Horst, G., Bonen, L., Tabak, H.F. and Grivell, L.A. (1986) *Cell* **44**, 225–234.
- Jacquier, A. (1990) *TIBS* **15**, 351–354.
- Lambowitz, A.M. and Perlman, P.S. (1990) *TIBS* **15**, 440–444.
- Lambowitz, A.M. and Belfort, M. (1993) *Ann. Rev. Biochem.* **62**, 587–622.
- Copertino, D.W. and Hallick, R.B. (1991) *EMBO J.* **10**, 433–442.
- Copertino, D.W., Shigeoka, S. and Hallick, R.B. (1992) *EMBO J.* **11**, 5041–5050.
- Meunier, B., Tian, G.L., Macadre, C., Slonimski, P.P. and Lazowska, J. (1990) in *Structure, Function and Biogenesis of Energy Transfer Systems*, Quagliariello, E., Papa, S., Palmieri, F. and Saccone, C. (eds.), Elsevier Science Publishers B.V., Amsterdam, pp. 169–174.
- Moran, J.V., Kennell, J.C., Zimmerly, S.J., Lambowitz, A.M., Butow, R.A. and Perlman, P.S., manuscript in preparation.
- Skelly, P.J., Hardy, C.M. and Clark-Walker, G.D. (1991) *Curr. Genet.* **20**, 115–120.
- Kennell, J.C., Moran, J.V., Perlman, P.S., Butow, R.A. and Lambowitz, A.M. (1993) *Cell* **73**, 133–146.
- Carignani, G., Groudinsky, O., Frezza, D., Schiavon, E., Bergantino, E. and Slonimski, P.P. (1983) *Cell* **35**, 733–742.
- Carignani, G., Netter, P., Bergantino, E. and Robineau, S. (1986) *Curr. Genet.* **11**, 55–63.
- Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A. and Perlman, P.S. (1993), manuscript in preparation.
- Michel, F. and Lang, B.F. (1985) *Nature* **316**, 641–643.
- Doolittle, R.F., Feng, D.-F., Johnson, M.S. and McClure, M.A. (1989) *Quart. Rev. Biol.* **64**, 1–30.
- Xiong, Y. and Eickbush, T.H. (1990) *EMBO J.* **9**, 3353–3362.
- McClure, M.A. (1991) *Mol. Biol. Evol.* **8**, 835–856.
- Bergantino, E. and Carignani, G. (1990) *Mol. Gen. Genet.* **223**, 249–257.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
- Hallick, R.B., Hong, L., Drager, R.G., Favreau, M.R., Monfort, A., Orsat, B., Spielmann, A. and Stutz, E. (1993) *Nucleic Acids Res.* **21**, 3537–3544.
- Montandon, P.-E., Vasserot, A. and Stutz, E. (1986) *Curr. Genet.* **11**, 35–39.
- Nargang, F.E., Bell, J.B., Stohl, L.L. and Lambowitz, A.M. (1984) *Cell* **38**, 441–453.
- Boer, P.H. and Gray, M.W. (1988) *EMBO J.* **7**, 3501–3508.
- Schuster, W. and Brennicke, A. (1987) *EMBO J.* **6**, 2857–2863.
- Steitz, T.A. (1993) *Current Opinion Struct. Biol.* **3**, 31–38.
- Siemeister, G., Buchholz, C. and Hachtel, W. (1990) *Mol. Gen. Genet.* **220**, 425–432.
- Johnson, M.S., McClure, M.A., Feng, D.-F., Gray, J. and Doolittle, R.F. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7648–7652.
- Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A. and Steitz, T.A. (1992) *Science* **256**, 1783–1790.
- Lang, F.B., Ahne, F. and Bonen, L. (1985) *J. Mol. Biol.* **184**, 353–366.
- Field, D.J., Sommerfield, A., Saville, B.J. and Collins, R.A. (1989) *Nucleic Acids Res.* **22**, 9087–9099.
- Neuhaus, H. and Link, G. (1987) *Curr. Genet.* **11**, 251–257.
- Rost, B. and Sander, C. (1992) *Nature* **360**, 540.
- Koonin, E.V., Boyko, V.P. and Dolja, V.V. (1991) *Virology* **181**, 395–398.
- Jozwik, C.E. and Miller, E.S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5053–5057.
- Mattaj, I.W. (1993) *Cell* **73**, 837–840.
- Swofford, D.L. (1990) *PAUP: Phylogenetic Analysis Using Parsimony; Version 3.0*. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.
- Jarrell, K.A., Dietrich, R.C. and Perlman, P.S. (1988) *Mol. Cell. Biol.* **8**, 2361–2366.
- Hebbar, S.K., Belcher, S.M. and Perlman, P.S. (1992) *Nucleic Acids Res.* **20**, 1747–1754.
- Lambowitz, A.M. (1989) *Cell* **56**, 323–326.
- Wolfe, K.H., Morden, C.W. and Palmer, J.D. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10648–10652.
- Higgins, D.G. and Sharp, P.M. (1989) *CABIOS* **5**, 151–153.
- Higgins, D.G. and Sharp, P.M. (1988) *Gene* **73**, 237–244.
- Lampson, B.C., Sun, J., Hsu, M.-Y., Vallejo-Ramirez, J., Inouye, S. and Inouye, M. (1989) *Science* **243**, 1033–1038.
- Sugita, M., Shinozaki, K. and Sugiura, M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3557–3561.
- Du Jardin, P. (1992) *Z11741*.
- Kolosov, V.L. and Zolotarev, A.S. (1990) *Bioorg. Khim.* **16**, 1060–1068.
- Boyer, S.K. and Mullet, J.E. (1988) *Nucleic Acids Res.* **16**, 8184.
- Sexton, T.B., Jones, J.T. and Mullet, J.E. (1990) *Curr. Genet.* **17**, 445–454.
- Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.R., Meng, B.Y., Li, Y.Q., Kanno, A., Nishizawa, Y., Hirai, A., Shinozaki, K. and Sugiura, M. (1989) *Mol. Gen. Genet.* **217**, 185–194.
- Lidholm, J. and Gustafsson, P. (1991) *Nucleic Acids Res.* **19**, 2881–2887.
- Tsudzuki, J., Nakashima, K., Tsudzuki, T., Hiratsuka, J., Shibata, M., Wakasugi, T. and Sugiura, M. (1992) *Mol. Gen. Genet.* **232**, 206–214.
- Umesono, K., Inokuchi, H., Shiki, Y., Takeuchi, M., Chang, Z., Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K. and Ozeki, H. (1988) *J. Mol. Biol.* **203**, 299–331.
- Kück, U. (1989) *Mol. Gen. Genet.* **218**, 257–265.
- Kono, M., Satoh, H., Okabe, Y., Abe, Y., Nakayama, K. and Okada, M. (1991) *Plant Mol. Biol.* **17**, 505–508.
- Chapdelaine, Y. and Bonen, L. (1991) *Cell* **65**, 465–472.
- Wissinger, B., Schuster, W. and Brennicke, A. (1991) *Cell* **65**, 473–482.
- Wahleithner, J.A., MacFarlane, J.L. and Wolstenholme, D.R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 548–552.
- Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., Kohchi, T. and Ohyama, K. (1992) *J. Mol. Biol.* **223**, 1–7.
- Osiewacz, H.D. and Esser, K. (1984) *Curr. Genet.* **8**, 299–305.
- Matsuura, E.T., Domenico, J.M. and Cummings, D.J. (1986) *Curr. Genet.* **10**, 915–922.
- Cummings, D.J., Michel, F., Domenico, J.M. and McNally, K.L. (1990) *J. Mol. Biol.* **212**, 269–286.
- Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E., Tzagoloff, A. and Macino, G. (1980) *J. Biol. Chem.* **255**, 11927–11941.
- Hardy, C.M. and Clark-Walker, G.D. (1991) *Curr. Genet.* **20**, 99–114.